

Stochastic model for the vocabulary growth in natural languages

Martin Gerlach¹ and Eduardo G. Altmann¹

¹*Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany*

We propose a stochastic model for the number of different words in a given database which incorporates the dependence on the database size and historical changes. The main feature of our model is the existence of two different classes of words: (i) a finite number of core-words which have higher frequency and do not affect the probability of a new word to be used; and (ii) the remaining virtually infinite number of noncore-words which have lower frequency and once used reduce the probability of a new word to be used in the future. Our model relies on a careful analysis of the google-ngram database of books published in the last centuries and its main consequence is the generalization of Zipf's and Heaps' law to two scaling regimes. We confirm that these generalizations yield the best simple description of the data among generic descriptive models and that the two free parameters depend only on the language but not on the database. From the point of view of our model the main change on historical time scales is the composition of the specific words included in the finite list of core-words, which we observe to decay exponentially in time with a rate of approximately 30 words per year.

PACS numbers: 89.65.-s, 89.75.Da, 87.23.Ge, 05.10.-a

I. INTRODUCTION

The finitude of the vocabulary size is an intuitive idea. It finds a definite argument in the fact that words cannot be arbitrary large and are built from the combination of a finite number of phonemes or letters. On the other hand, even in our time of big data [1–3] there is no indication of a saturation of the vocabulary size (total number of different words) with increasing database size. In order to clarify whether the finitude of the vocabulary has any practical consequence, it is essential to understand not only the birth and death of words [4–6], but also how the number of different words depends on the database size. The interest in this problem is motivated by linguistic studies [7, 8] as well as by applications in search engines, which require an estimation of the number of different words expected in a given database [9–11].

The scaling between the number of different words, N , and the size of the database in words, M , as $N \sim M^\lambda$ is known as Heaps' law [12] and has been studied in different linguistic [13–15] and non-linguistic [16, 17] contexts. The universality and interest of this empirical scaling is surpassed only by Zipf's law [18], which states that the frequency $F(r)$ of the r -th most frequent word in a database decays as $F(r) \sim 1/r$. The relation between Heaps' and Zipf's law has been the subject of great recent interest [19–21]. Furthermore, it is well known that deviations of the Heaps'- and Zipf's-laws are observed in the tails of the Heaps'- and Zipf's- plots (i.e., for large N and r , respectively) [22–24]. Similar deviations of fat-tailed distributions appear in a variety of social and physical systems [25, 26] and are crucial when extrapolating to the limit of large databases.

In this paper we propose a stochastic growth model whose predictions go beyond the simpler scalings of Heaps' and Zipf's law and are compatible with actual observations in the tail of the corresponding distributions. Our model is in the same spirit of, but differs from, the

simpler versions of Yule's-, Simon's-, Gibrat's-, and preferential attachment- growth models [25, 27], because it contains two categories of words and leads to two scaling regimes in the Heaps'- and Zipf's- plots. Our statistical analysis in the google-ngram database indicates that the only two free parameters of our model remain unchanged over centuries, depend only on the language, and that there is a slow change of words belonging to each category. The latter adds to the recent interest in language dynamics as a complex system [28, 29].

The paper is organized as follows: in Sec. II we present statistical analysis of the google-ngram database in terms of word frequencies as well as the growth of the vocabulary. This will then lead us to the formulation of our stochastic model for the vocabulary growth in Sec. III. In Sec. IV we investigate dynamical aspects on historical time scales within the framework of our model.

II. DATA ANALYSIS

Data

The main motivation for our model comes from empirical observations. As databases, we use the google-ngram corpus [1] for English, German, French, Spanish, and Russian, which provide data of the word-frequencies (occurring in printed books) with a yearly resolution for a period of several hundred years (1520-2000). Our main interest in this database stems from its large size (several millions of books with $> 10^{11}$ words) and from the long time span it covers (thus enabling us to trace historical changes in the usage of language). We consider as words only the 1-grams consisting uniquely by letters present in the alphabet of the corresponding language. This pragmatic definition guarantees that our observations are not affected by symbolic sequences, foreign words, numbers, or scanning problems. For each language we use two dif-

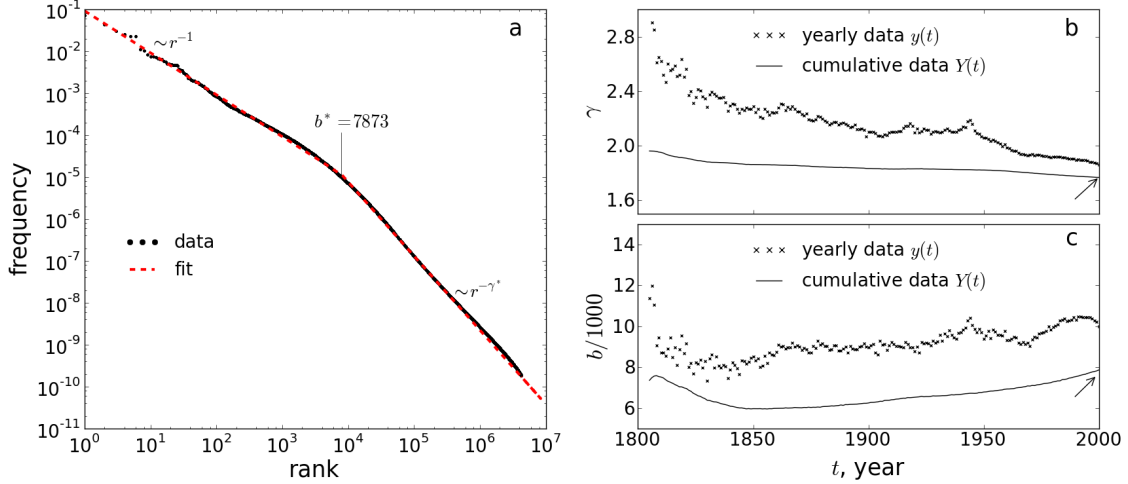


FIG. 1. Rank-frequency distribution shows double scaling behavior (Zipf’s plot). a) Rank-frequency distribution for the English database $Y(2000)$ (solid) and a ML-fit of Eq. (1) (dashed). b+c) parameters γ and b obtained from ML-fits of Eq. (1) to yearly $y(t)$ (x-symbols) and accumulated $Y(t)$ (solid) database for $t \in [1805, 2000]$. Arrows indicate the values of the parameters γ^* and b^* obtained for the fit in a).

ferent partitions of the database: i) yearly (y), in which case $y(t)$ corresponds to the database of the year t ; and ii) cumulative (Y), in which case $Y(t) = \sum_{t_0}^t y(t)$. We consider only words which appeared at least $n = 41$ times in order to avoid biases due to the filtering mechanism used in the google-ngram database, see Supplementary Information (SI) Sec. I for further details. Here we show our detailed analysis for the largest database (English, $N = 335$, $t_0 = 1520$, consistent data for $t \in [1805, 2000]$). For the other 4 languages we report the main findings and leave the details for the SI.

Zipf’s analysis

Our first empirical analysis focus on the distribution of word frequencies. In his seminal work, Zipf proposed that the frequency of the r -th most frequent word in a given text is given by $F(r) = F(1)/r$ [18]. It is easy to see that this scaling has to break for large r : due to the divergence of the harmonic series for sufficiently large databases one arrives at $\sum_{r=1}^N F(r) > 1$ (sum of frequencies larger than text size). In English $F(1) \approx 0.07$ (the frequency of “the”) and $\sum_{r=1}^N F(r) > 1$ for $N \approx 10^6$, meaning that $F(r)$ has to decay faster than $1/r$ for $r \gtrsim 10^6$. This well-known expectation, which is clearly seen in our data shown in Fig. 1(a), motivated numerous different generalization of Zipf’s proposal [30–32]. While many of these proposals were shown to provide a better account of particular databases, they remain in a great extent unsatisfactory because they lack the simplicity and universality of Zipf’s original proposal (e.g., the parameters vary depending on the size, topic or date of publication of the analyzed texts [33, 34]).

Motivated by the new magnitude of our large database, we apply rigorous statistical tests to determine which of the previously proposed distributions provide a better account of the data. We select 7 of the most popular previously-proposed heavy-tailed distributions with at most 2 free parameters [8, 23, 24]: power-law, two power-laws, shifted power-law, log-normal, Weibull, and power-laws with exponential cutoffs (in the tail and beginning, respectively). The parameters for each distribution were obtained numerically by means of Maximum Likelihood (ML) estimation [35]. In addition we i) calculate the probability that the data was generated by that model (χ^2 p -value [36]) and ii) compare which model is more likely to describe the data (relative likelihood [37]) for each fit (for details see SI-Sec. II A).

The results show that it is extremely unlikely ($P < 10^{-15}$) that the data was drawn exactly from any of the proposed distributions, a consequence of the large databases which makes any small (true) deviation incompatible with these simple fits. On the other hand, the results show unequivocally that for English the distribution with two power-laws is clearly the best fit ($1 - P < 10^{-15}$) for all databases with a size of $> 10^9$ tokens.

We now discuss in detail the best two-parameter model we identify from our data:

$$F_{dp}(r; \gamma, b) = C \begin{cases} r^{-1}, & r \leq b \\ b^{\gamma-1} r^{-\gamma}, & r > b, \end{cases} \quad (1)$$

where b , and γ are free-parameters and $C = C(\gamma, b)$ is the normalization constant. The effect of the threshold n applied to the frequency of words is that, in practice, data of $F(r)$ is limited to $F(r) \geq n/M$ (M is the observed number of tokens). The original Zipf’s law is recovered

for high-frequency words and a critical rank $r = b$ determines a transition to a power-law with exponent γ . Double power-laws were proposed as a generalization of Zipf's law in Ref. [38] and further investigated in Refs. [39, 40]. These insightful works used distributions with two power-law exponents γ_1, γ_2 and were motivated by the visual inspection of double logarithmic plots. Our improved statistical analysis confirm and extend these observations for the simpler distribution Eq. (1). Besides the likelihood analysis and visual inspection given in Fig. 1, a third strong evidence in favor of distribution (1) comes from the comparison of the estimated parameters of different corpora shown in Fig. 1(b,c). Very similar values $b \in [7 \cdot 10^3, 12 \cdot 10^3]$ and $\gamma \in [1.8, 2.5]$ were obtained for non-overlapping databases, and the fluctuations become smaller for increasing database size. The observations strongly suggest that the same fixed parameters provide a good description of all English texts (e.g., $y(1900)$ and $y(2000)$). In order to test this, hereafter we do not consider individual fits for each database and instead assume that Eq. (1) is valid with $b = b^* = 7873$ and $\gamma = \gamma^* = 1.77$, values obtained for our largest database $Y(2000)$.

Similar findings also apply to the other languages. In Tab. I we summarize the parameters γ^* and b^* obtained from a ML-fit of the largest database $Y(2000)$ of the respective language to Eq. (1). French and Spanish are also best described by Eq. (1) for databases exceeding a particular size and yield values for γ^* and b^* similar to English. For German and Russian Eq. (1) constitutes only the second best model. However, we have strong indications that it provides a better account of the tails ($r \gg b^*$) and therefore we expect that even larger databases will reveal the double power-law as the best fit also in these languages (see SI-Sec. II B for details). Apart from being the smallest databases among the investigated languages, another feature affecting the fitting in German and especially in Russian is the higher degree of inflection in the morphology of these languages. We recall that no lemmatization was applied in our definition of words and, therefore, inflected words (obtained, e.g., by adding a suffix) are counted as new word-types. This reasoning explains the higher measured values of b^* (vocabulary in the r^{-1} regime). From the fitting perspective, however, the large values of b^* in German and Russian require even larger databases to characterize the deviations from the r^{-1} regime for $r \gg b^*$.

language	b^*	γ^*
English	7,873	1.77
French	8,208	1.78
Spanish	8,757	1.78
German	19,863	1.62
Russian	62,238	1.94

TABLE I. Parameters b^* and γ^* obtained from ML-fit of Eq. (1) obtained for the largest database $Y(2000)$ for all considered languages.

Heaps' analysis

We now turn to our second empirical analysis: the dependence of the total number of different words (word-types, N) on the size of the database (in word-tokens, M). The classical result for this relation is the empirical Heaps' law [12], which states that $N \sim M^\lambda$ with $\lambda \in [0, 1]$ ($a \sim b$ indicates that $a/b = \text{constant}$ for large b). We start searching for the consequences of our previous observations in the Zipf's analysis to this new problem. A simple and powerful approach is the so-called Zipfian ensemble (ZE) [21], which assumes that the occurrence of every possible word is governed by a Poisson process with an intensity proportional to its frequency (see SI-Sec. III A). It was shown that under this or similar assumptions (e.g., stochastic processes with fixed frequencies for words), asymptotically Heaps' law can be interpreted as a direct consequence of a Zipfian rank frequency distribution $F(r) \sim r^{-\gamma}$ [9, 13, 14, 19, 21] and vice versa [20, 41, 42], where $\gamma = 1/\lambda$. Here we want to draw attention to the fact that these observations are not restricted to Zipf's and Heaps' laws, i.e., assuming a stochastic model, the relationship between $F(r)$ and $N(M)$ can always be established. The expectation of the ZE of Eq. (1) with a threshold $n \gg 1$ is (see SI-Sec. III B)

$$N_{\text{dp}}(M; \gamma, b) = C_n \begin{cases} M, & M \ll M_b \\ M_b^{1-1/\gamma} M^{1/\gamma}, & M \gg M_b, \end{cases} \quad (2)$$

where M_b is the number of tokens such that $N(M_b) = b$ and the scaling constant $C_n = C/n$ ($C \approx F(1)$ being the frequency of the most common word, as can be seen from Eq. (1)). Thus, the effect of the threshold n applied to the growth curve of the vocabulary simply amounts to rescaling the constant C . While the expected (average) number of word-types over many realizations of the stochastic process leads to a sharp transition between the two regimes, the values of $N_{\text{dp}}(M \approx M_b)$ might depend more strongly on the particular realization.

In Fig. 2 we show that the data in the google-ngram database obeys the scalings of Eq. (2). In Fig. 2(a) we present the $N(M)$ curve for English. While for the yearly database $y(t)$ we obtain a set of points for each t , the cumulative database $Y(t)$ builds a curve of vocabulary growth for increasing t . Despite the differences in these databases, all the data lie in a relatively narrow region of the plot which resembles a single curve compatible with the double scaling of Eq. (2). This curve is well described by the $N(M)$ curve obtained from the combination of the double power-law distribution Eq. (1) with fixed parameters (γ^*, b^*) and the assumption of Poisson usage of words, in the spirit of the ZE. Similar observations apply to all considered languages, as shown in Fig. 2(b). On closer inspection, Fig. 2(c), the fine details of the $N(M)$ curve are not compatible with the fluctuations expected from the strongly simplifying assumptions of the ZE. It is, nevertheless, remarkable that the agree-

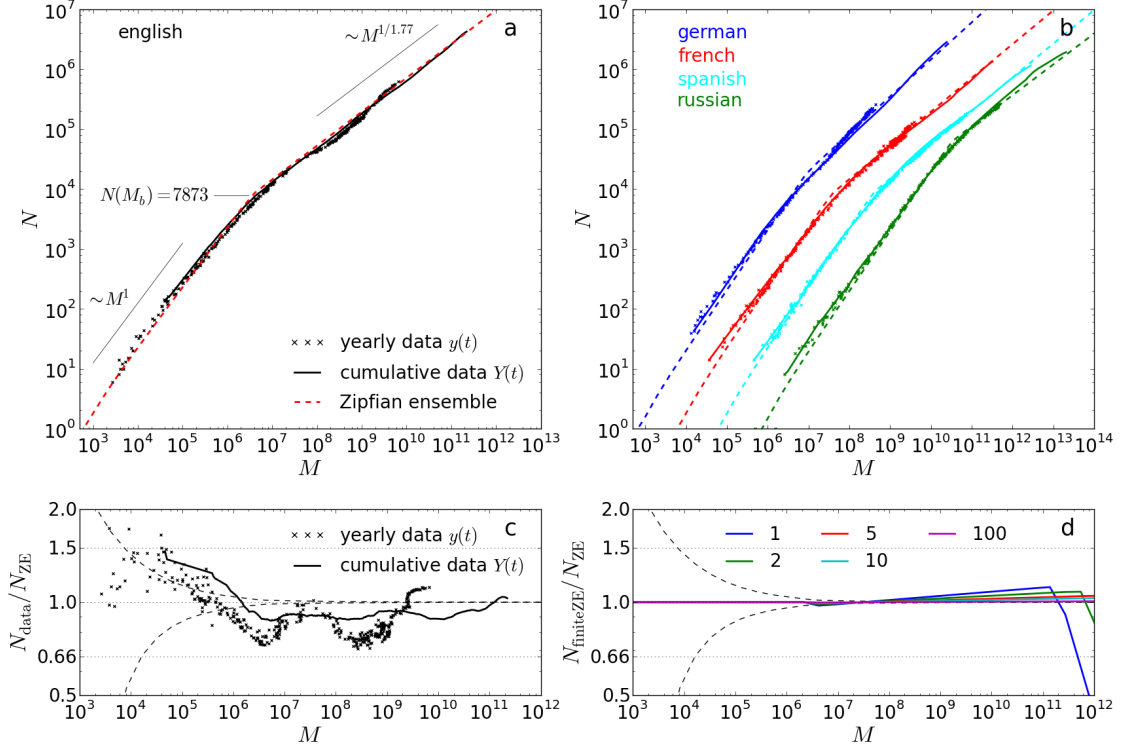


FIG. 2. Vocabulary N as a function of database size M (Heaps' plot). a) Number of word-types as a function of word-tokens for yearly $y(t)$ (x-symbols) database, cumulative $Y(t)$ (solid) database, and the Zipfian ensemble (dashed) assuming $n = 41$ and the rank-frequency distribution Eq. (1) with $b^* = 7873$ and $\gamma^* = 1.77$. b) Same curves as in a) but for different languages showing the same scaling behaviour. In order to increase visibility the curves for French, Spanish, and Russian were shifted, respectively, by one, two, and three decades with respect to their x-values. c) Deviation of the data $y(t)$ and $Y(t)$ (N_{data}) from the ZE growth curve (N_{ZE}). The dashed lines show the 95%-confidence interval of the ZE. d) Deviation of a ZE growth curve with a hypothetically finite vocabulary (N_{finiteZE}) from the ZE growth curve with infinite vocabulary (N_{ZE}) assuming rank-frequency distribution Eq. (1). Possible size of total vocabulary is given in units k of the number of observed types in $Y(2000)$, such that $N_{\text{ZE}}^{\text{max}} = k \cdot 4,263,717$ with $k = 1, 2, 5, 10, 100$. Since for $M \rightarrow \infty$: $N_{\text{finiteZE}}(M) \rightarrow N_{\text{ZE}}^{\text{max}}$ the deviation for $k = 1$ becomes already large for $M > 10^{11}$.

ment between model and data remains within 50% for different databases and over 9 orders of magnitude in size.

Here it is worth revisiting the question about the finitude of the vocabulary. Even after more than 10^6 different words the $N(M)$ data in Fig. 2 does not seem to be saturated. To further investigate this point, we perform the ZE with the same rank-frequency distribution from Eq. (1) (fixed b^* , γ^*) but varying the maximum possible number of different words $N_{\text{ZE}}^{\text{max}}$, e.g., 1, 2, 5, 10, and 100 times the observed number of distinct words in our largest database $Y(2000)$. It can be seen in Fig. 2(d) that the differences for the predicted growth curves for such different hypothetical vocabulary sizes are negligible compared to the fluctuations of the real data. From this we conclude that given the data accessible so far the possible vocabulary can be regarded for all practical purposes to be infinite (although bounded by combinatorial arguments due to a finite alphabet and word length). The fact that the same distribution Eq. (1) with fixed

parameters accounts for the observation across all years shows that the observation of different number of words is driven mainly by the different database size and not by a change in vocabulary richness over time.

III. MODEL

In this section we propose a simple generative model which recovers and allows for an improved interpretation of the double scalings in our empirical findings – Eqs. (1) and (2). In line with the tradition of stochastic growth models explaining fat-tailed distributions, we consider an extension of Yule's model [25] which contains two classes of word-types: a core vocabulary and a non-core vocabulary [39]. At each step a word-token is drawn ($M \mapsto M + 1$) and attributed to a word-type depending on probabilities specified below, see Fig. 3 for a sketch of the model. The total number of different word-types is given by $N = N_c + N_{\bar{c}}$, where ($N_{\bar{c}}$) N_c is the number

of (non)core-words. The new word-token can either be a new word-type ($N \mapsto N + 1$) with a probability p_{new} or an already existing type ($N \mapsto N$) with probability $1 - p_{\text{new}}$. In the latter case, a (previously used) type is attributed to the token at random with probability proportional to the number of times this type has occurred before. In the former case, the new type can either originate from a finite set of N_c^{max} core-words ($N_c \mapsto N_c + 1$) with probability p_c or come from a potentially infinite set of noncore-words ($N_{\bar{c}} \mapsto N_{\bar{c}} + 1$). In our simplest model we consider p_c to be a constant which becomes zero only if all core-words were drawn ($N_c = N_c^{\text{max}}$):

$$p_c(N_c) = \begin{cases} p_c^0 & \text{if } N_c < N_c^{\text{max}}, \\ 0 & \text{if } N_c = N_c^{\text{max}}. \end{cases} \quad (3)$$

The final element of our model, which establishes the distinguishing aspect of core-words, is the dependence of p_{new} on N . By definition, we think of core-words as belonging to the central vocabulary of the language and, therefore, the usage of a new core-word does not affect the probability of using a new word-type in the future, i.e., $p_{\text{new}} = p_{\text{new}}(N_{\bar{c}})$. On the other hand, if a noncore-word is drawn ($N_{\bar{c}} \mapsto N_{\bar{c}} + 1$) the p_{new} for future choices should decrease. We saw that there is no indication of an upper bound on the vocabulary ($p_{\text{new}} > 0$) over many decades, which suggests that the reduction of p_{new} with $N_{\bar{c}}$ is slow. Accordingly, we assume

$$p_{\text{new}} \mapsto p_{\text{new}} \left(1 - \frac{\alpha}{N_{\bar{c}} + s} \right), \quad (4)$$

with the decay rate $\alpha > 0$ and the constant $s \gg 1$ which is introduced simply in order to damp the reduction of p_{new} for small $N_{\bar{c}}$ (for simplicity, we use $s = N_c^{\text{max}}$).

We now show how this model recovers Eqs. (1) and (2). We require that $1 - p_c^0 \ll 1$, which simply means that it is much more likely to draw core-words than noncore-words initially. In this case we can obtain approximately exact solutions for $N(M)$ in the two limiting cases considered in Eq. (2). When $N_c \ll N_c^{\text{max}}$, $1 - p_c \ll 1$ so that $p_{\text{new}} \approx \text{const.}$ and therefore we trivially obtain that $N \sim M^\lambda$ with $\lambda \approx 1$. This case resembles the very beginning of the vocabulary growth, when most new word-types belong to the set of core-words. In the case $N_{\bar{c}} \gg N_c^{\text{max}}$, $p_c = 0$ and $N \approx N_{\bar{c}}$ so that Eq. (4) becomes in the continuum limit:

$$\frac{d}{dN} p_{\text{new}}(N) = -\alpha \frac{p_{\text{new}}(N)}{N}, \quad (5)$$

from which it follows that $p_{\text{new}} \sim N^{-\alpha}$. Thus we see that Eq. (4) is a minimal assumption that guarantees that the vocabulary can in principle be of infinite size.

We now obtain the expected growth curve $N(M)$. Notice that our model can be considered a biased random walk in N , which, as an approximation, can be mapped onto a binomial random walk by the coordinate transformation $N(M)$ such that $p_{\text{new}}(N) = p_{\text{new}}(N(M))$. The

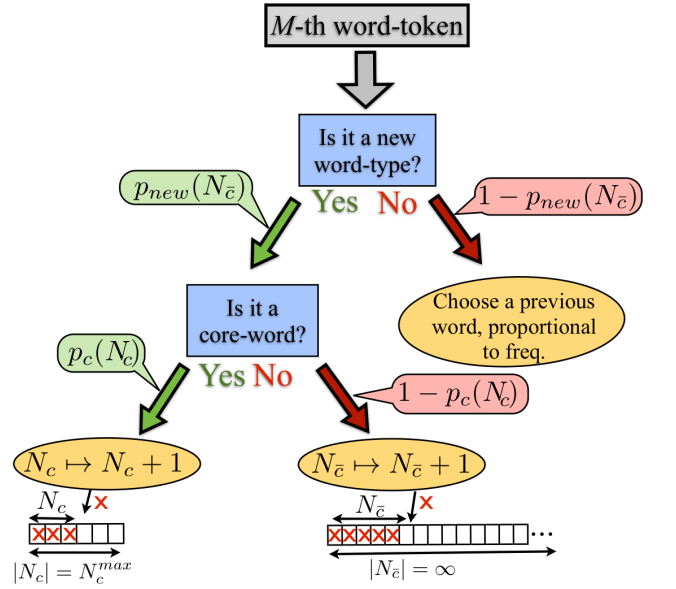


FIG. 3. Illustration of our generative model for the usage of new words.

resulting Poisson-Binomial process [43] can be treated analytically, e.g., the transformation $N(M)$ is then given by the average of the vocabulary growth:

$$\begin{aligned} N(M) &= \int_1^M dM' p_{\text{new}}(M') \\ &= \int_1^{N(M)} dN' \left| \frac{dM'}{dN'} \right| p_{\text{new}}(N'). \end{aligned} \quad (6)$$

Using $p_{\text{new}} \sim N^{-\alpha}$, this equation holds (self-consistently) only by assuming a sub-linear growth for the vocabulary $N \sim M^\lambda$, where the relation $\lambda = (1 + \alpha)^{-1}$ is established. In accordance with Eq. (2), we identify the following relation between the parameters: $N_c^{\text{max}} = b$ and $\alpha = \gamma - 1$. The fitting parameters of Eq. (1) can thus be interpreted as: b is the size of the core vocabulary and γ controls the sensitivity of the probability of using a new word to the number of already used words in Eq. (5).

Since the probability of usage for already used word-types is assumed to be proportional to the number of times it occurred before, we guarantee that Eq. (2) implies (1) [20], meaning that the double scaling in the Zipf plot is also recovered from our generative model. While the previous arguments show that the correct scalings are obtained by our model, in order to obtain an agreement with the data it is essential to: (i) use the normalization constant C in order to determine the initial probability of finding a new word in Eq. (4); (ii) re-scale the distribution using the threshold n as M/n ; and (iii) account for the disproportionately large weight of the first word types (in the Zipf plot). Taking these points into account, direct simulations of the model in Fig. 3 with the traditional parameters $b = b^*$ and $\gamma = \gamma^*$ lead to Zipf's and Heaps' curves, which resemble the original fits. See SI-Sec. IV for all details.

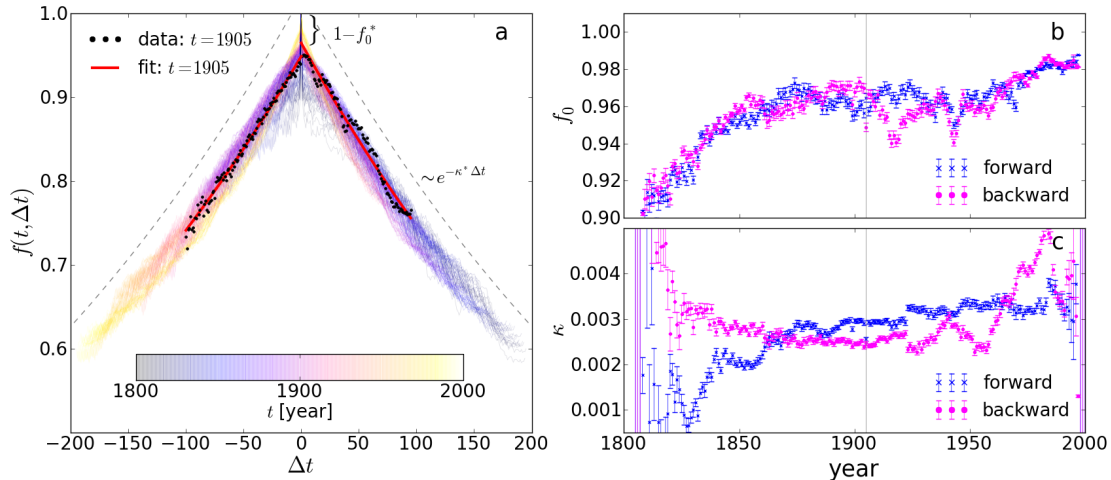


FIG. 4. Historical change in the composition of core-words in the English vocabulary. a) fraction $f(t, \Delta t)$ of core-words in $y(t)$ that remain among this set in $y(t + \Delta t)$ for $t \in [1805, 2000]$ (pale colors) and in particular for $t = 1905$ (black dots) with the corresponding exponential fit (red line). b+c) Parameters f_0 and κ in the exponential decay Eq. (7) of the curves in a) obtained through least-square fits. Forward (backward) decay refers to $\Delta t > 0$ ($\Delta t < 0$).

IV. HISTORICAL CHANGES

The model described so far has been shown to give a good account for all databases and all years with the same fixed two parameters $N_c^{\max} = b^* = 7,873$ and $\alpha = \gamma^* - 1 = 0.77$ in the case of English. A natural question is, therefore, what actually changes in historical time scales? Considering two different databases (say two different years), our model does not consider any differences in the actual composition of the database. Even if the value of N_c^{\max} remains constant this does not mean that the *same* word-types are observed for all years. From the point of view of our model, the main change a word-type can experience is to enter or to leave the group of core-words.

Following this thought, we investigate the composition of the set of core-words in the yearly databases $y(t)$ in the time $t \in [1805, 2000]$ in Fig. 4. We calculate the fraction $f(t, \Delta t)$ of core-words from $y(t)$ that remains among this set in $y(t + \Delta t)$. Figure 4(a) shows that all curves can be qualitatively described by an exponential decay

$$f(t, \Delta t) = f_0 e^{-\kappa |\Delta t|}, \quad (7)$$

independent of whether forward ($\Delta t > 0$) or backward time ($\Delta t < 0$) was considered. This is further supported in Fig. 4(b-c), where the parameters f_0 and κ obtained numerically from a least-square fit [35] of Eq. (7) for all curves $f(t, \Delta t)$ with $t \in [1805, 2000]$ are presented. In order to avoid biases due to different number of points in the fit, for each t we performed a fit with the same number of points $\min\{2000 - t, t - 1805\}$ forwards and backwards in time. On closer inspection, two features connected to the interpretation of the parameters f_0 and κ deserve a more careful discussion. The parameter $f_0 < 1$

represents the discontinuous change of core-words in two subsequent years. It strongly depends on the different selection of books in the construction of the respective databases and can be attributed to the finite size of the database, which leads to a wrong estimation of the “true” core-words. Consistently with this interpretation, Fig. 4(b) shows that f_0 becomes smaller over time, due to the fact that database size increases leading to a better sampling of words. Nevertheless, a value of $f_0 \approx 0.98$ indicates that this is still far from being negligible (e.g., for $N_c^{\max} = 7,873$ this means that around 150 words of the set of core-words will be different due to finite sampling). In contrast, the decay rate κ describes the continuous replacement of core-words over time. The most intriguing observation in Fig. 4(c) is that this change experiences an acceleration over time as κ grows by more than 50% from 1805 to 2000.

V. DISCUSSION

In summary, we have shown that the rank frequency distribution and the vocabulary growth of languages can be best described by simple two-scaling functions. The only two free parameters of the functions are related to each other and remain almost unchanged over centuries as well as databases and depend only on the considered language. We have also shown that these empirical findings can be interpreted as the result of a finite number of words belonging to a core vocabulary, which have different properties from the remaining virtually unlimited number of words, as summarized in Tab. II. This conclusion was achieved based on a simple generative stochastic model for vocabulary growth, which should be considered

as a null model for the prediction of fluctuations and vocabulary sizes. Finally, we found that the composition of the core-vocabulary experiences an exponential decay with a rate of 30 words per year, which is, remarkably, steadily accelerating in the past decades.

	Core Words	Non-core-words
Number	finite: $ N_c \in [10^3, 10^4]$	infinite: $ N_{\bar{c}} \rightarrow \infty$
Frequency	larger ($r > b^*$)	smaller ($r < b^*$)
Effect on p_{new}	none	reduction

TABLE II. Properties of core (c) and noncore (\bar{c}) words in our model.

It is worth comparing these findings in view to previous results. As far as we are aware, our analysis provides the first rigorous statistical confirmation of similar previous proposals [38–40] of the double-scaling generalizations of Zipf’s law – Eq. (1). The consequence of this to vocabulary growth and Heap’s law (see also [40]), which we drew based on a Poisson usage of words [21], is that the rate of introduction of new words decays but never vanishes with increasing database size. This is in contrast to recent claims which reported a convergence to a maximum vocabulary size [14]. We note that this previous analysis was based on single books and therefore the database sizes were close to our transition point N_c^{max} , which we believe was misinterpreted with a systematic decay. A generalization of a Yule’s type process to obtain double-scaling degree distribution in a network of words was introduced in Ref. [44]. Two crucial differences to our model are that it yields fixed exponents and cannot be understood as a generative model of texts (token by token). Interestingly, in Ref. [6] an analysis of the network constructed from the thesaurus also showed the existence of a set of core-words almost of the same size as ours.

Our simple model and expression for the vocabulary growth as a function of database size has important practical consequences. Simply knowing the database size (in tokens, M , or potentially in bits), and using the language dependent parameters ($C, N_c^{\text{max}}, \alpha$) reported above, from Eq. (2) one can immediately estimate the expected number of different words, N , appearing more than n times. This is crucial for search engines and data mining programs because it allows for an estimation of the memory to be allocated prior to the scanning of an unknown database, e.g., in the construction of the inverted index [9–11]. Even the fluctuations around this expectation can be easily computed through our generative model or through the Poisson assumption of word-usage. Of course, this strong assumption ignores correlations and typically underestimates the expected fluctuations, so that our model should be considered as the simplest null model. The existence of a transition between two scalings (which is under the reach of even single large books) shows that simple estimations based only on the traditional Zipf’s law have to be generalized. For instance, a

commonly used index of vocabulary richness of a text is Herdan’s coefficient given by the ratio $\log N / \log M$ [8]. In view of our results, the coefficient is highly dependent on which of the two scaling regimes is reached with the given size of the text.

We now compare our observations of change on historical time scales to other historical changes in language usage. For the whole vocabulary, we obtain that its size is mainly driven by the available database size. This is in contrast to previous conclusions based on the same google-ngram database which detected a growth of vocabulary in time [1]. Here it is important to note that this previous analysis included a substantially different filtering of the listed 1-grams to achieve valid words in the vocabulary, including a frequency criterion and manual classification. Still, our results show that also in this case a more careful analysis of the role of the database size is needed. For the core vocabulary, we observe a fairly constant number of constituents over centuries. The number of words common to core-vocabularies of different databases was found to decay exponentially with the time between publication of the databases, e.g., for English the decay rate is approximately 30 words per year and the half-life of a word in the core vocabulary is 200 years. A decay rate for regularization of verbs was reported in Ref. [5] with a half-life varying between 750 and more than 10,000 years as well as for a fundamental vocabulary of 200 words with a half-life varying between 300 and 38,800 years [4]. Perhaps our most intriguing finding is the approximately linear increase of the rate in time, which eventually confirms the overall acceleration of language change and society in general, as propagated in Ref. [1].

Our results can be extended in many directions and open new possibilities of studies of vocabulary change. Directly related to our observations and model, it remains to be explained the specific value of the parameter $\gamma^* \approx 1.77$, which is intriguingly similar across different languages. Another important point is to assess the limitations of our estimations due to the role of correlations inside real texts and databases, and how this could be introduced into our model. Furthermore, it remains to be shown whether the transition between two scalings due to the existence of a core vocabulary can be related to the phenomenon of phase transitions in ranking stability of complex systems recently reported in Ref. [45]. Finally, we believe that our model provides the correct null model for normalizations due to database sizes and that therefore future investigations of historical effects on the vocabulary should take this into account.

ACKNOWLEDGMENTS

We are indebted to J. Miotto, and R. Guimerà for valuable discussion about the data analysis. We thank F. Ghanbarnejad, and J. Leitao for careful reading of the manuscript.

-
- [1] J.-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden, *Quantitative analysis of culture using millions of digitized books*, Science **331**, 176 (2011).
 - [2] J. Gao, J. Hu, X. Mao, and M. Perc, *Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries* J. R. Soc. Interface **9**, 1956 (2012).
 - [3] A.M. Petersen, J. Tenenbaum, S. Havlin, and H.E. Stanley, *Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death*, Sci. Rep. **2**, 31 (2012).
 - [4] M. Pagel, Q.D. Atkinson, and A. Meade, *Frequency of word-use predicts rates of lexical evolution throughout Indo-European history*, Nature (London) **449**, 717 (2007).
 - [5] E. Lieberman, J.-P. Michel, J. Jackson, T. Tang, and M.A. Nowak, *Quantifying the evolutionary dynamics of language*, Nature (London) **449**, 713 (2007).
 - [6] D. Levary, J.-P. Eckmann, E. Moses, and T. Tlusty, *Loops and Self-Reference in the Construction of Dictionaries*, Phys. Rev. X **2**, 031018 (2012).
 - [7] G. Wimmer, and G. Altmann, *Review Article: On Vocabulary Richness*, J. Quant. Linguist. **6**, 1 (1999).
 - [8] R.H. Baayen, *Word Frequency Distributions* (Kluwer Academic Publishers, 2001).
 - [9] R. Baeza-Yates, and G. Navarro, *Block addressing indices for approximate text retrieval*, J. Am. Soc. Inform. Sci. **51**, 69 (2000).
 - [10] H.E. Williams, and J. Zobel, *Searchable words on the Web*, Int. J. Digit. Libr. **5**, 99 (2005).
 - [11] B. Croft, D. Metzler, and T. Strohmann, *Search Engines: Information Retrieval in Practice* (Addison-Wesley, 2009).
 - [12] H.S. Heaps, *Information Retrieval* (Academic Press, 1978).
 - [13] M.A. Serrano, A. Flammini, and F. Menczer, *Modeling statistical properties of written text*, Plos One **4**, e5372 (2009).
 - [14] S. Bernhardsson, L.E. Correa da Rocha, and P. Minnhagen, *The meta book and size-dependent properties of written language*, New J. Phys. **11**, 123015 (2009).
 - [15] Y. Sano, H. Takayasu, and M. Takayasu, *Zipf's Law and Heaps' Law Can Predict the Size of Potential Words*, Prog. Theor. Phys. Supp. **194**, 202 (2012).
 - [16] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto, *Collective dynamics of social annotation*, P. Natl. Acad. Sci. USA **106**, 10511 (2009).
 - [17] R.W. Benz, S.J. Swamidass, and P. Baldi, *Discovery of power-laws in chemical space*, J. Chem. Inf. Model. **48**, 1138 (2008).
 - [18] G.K. Zipf, *The Psycho-Biology of Language* (Routledge, 1936).
 - [19] D.C. van Leijenhorst, and T.P. van der Weide, *A formal derivation of Heaps' Law*, Inform. Sciences **170**, 263 (2005).
 - [20] D.H. Zanette, and M.A. Montemurro, *Dynamics of Text Generation with Realistic Zipf's Distribution*, J. Quant. Linguist. **12**, 29 (2005).
 - [21] I. Eliazar, *The growth statistics of Zipfian ensembles: Beyond Heaps' law*, Physica A **390**, 3189 (2011).
 - [22] M.A. Montemurro, *Beyond the Zipf-Mandelbrot law in quantitative linguistics*, Physica A **300**, 567 (2001).
 - [23] W. Li, P. Miramontes, and G. Cocho, *Fitting Ranked Linguistic Data with Two-Parameter Functions*, Entropy **12**, 1743 (2010).
 - [24] G. Jäger, *Power laws and other heavy-tailed distributions in linguistic typology*, Adv. Complex Syst. **15**, 1150019-1 (2012).
 - [25] M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemp. Phys. **46**, 323 (2005).
 - [26] M.P.H. Stumpf, and M.A. Porter, *Critical Truths About Power Laws*, Science **335**, 665 (2012).
 - [27] M. Mitzenmacher, *A Brief History of Generative Models for Power Law and Log-normal Distributions*, Internet Math. **1**, 226 (2004).
 - [28] C. Castellano, S. Fortunato, and V. Loreto, *Statistical physics of social dynamics*, Rev. Mod. Phys. **81**, 591 (2009).
 - [29] A. Baronchelli, V. Loreto, and F. Tria, *Language Dynamics*, Adv. Complex Syst. **15**, 1203002-1 (2012).
 - [30] B. Mandelbrot, *An informational theory of the statistical structure of language*, Commun. Theor. , 486 (1953).
 - [31] J. Tuldava, *The frequency spectrum of text and vocabulary*, J. Quant. Linguist. **3**, 38 (1996).
 - [32] S.K. Baek, S. Bernhardsson, and P. Minnhagen, *Zipf's law unzipped*, New J. Phys. **13**, 043004 (2011).
 - [33] A. Cohen, R.N. Mantegna, and S. Havlin, *Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts*, Fractals **05**, 95 (1997).
 - [34] R. Ferrer i Cancho, *The variation of Zipf's law in human language*, Eur. Phys. J. B **44**, 249 (2005).
 - [35] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes* (Cambridge University Press, 2007), 3rd ed.
 - [36] R.B. D'Agostino, *Goodness-of-Fit-Techniques* (Taylor Francis, 1986).
 - [37] H. Akaike, *A new look at the statistical model identification*, IEEE T. Automat. Contr. **19**, 716 (1974).
 - [38] S. Naranan, and V. Balasubrahmanyam, *Models for power law relations in linguistics and information science*, J. Quant. Linguist. **5**, 35 (1998).
 - [39] R. Ferrer i Cancho, and R.V. Solé, *Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited*, J. Quant. Linguist. **8**, 165 (2001).
 - [40] A.M. Petersen, J. Tenenbaum, S. Havlin, and H.E. Stanley, *Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death*, Arxiv:1107.3707v1 , 31 (2011).
 - [41] H.A. Simon, *On a Class of Skew Distribution Functions*, Biometrika **42**, 425 (1955).
 - [42] A.P. Masucci, A. Kalampokis, V.M. Eguíluz, and E. Hernández-García, *Wikipedia information flow analysis reveals the scale-free architecture of the semantic space*, Plos One **6**, e17333 (2011).
 - [43] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume I*, (Wiley, 1968), 3rd ed.
 - [44] S.N. Dorogovtsev, and J.F.F. Mendes, *Language as an evolving word web*, Proc. R. Soc. Lond. B Biol. Sci. **268**, 2603 (2001).

- [45] N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.-P. Bouchaud, and A.-L. Barabási, *Dynamics of Ranking Processes in Complex Systems*, Phys. Rev. Lett. **109**, 128701 (2012).

Supplementary Information for: Stochastic model for the vocabulary growth of natural languages

Martin Gerlach¹ and Eduardo G. Altmann¹

¹ *Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany*

I. DATA

The data obtained from the google-ngram database [1] is filtered in two steps. First, we decapitalize each word (e.g. 'the' and 'The' are counted as the same word) and further restrict ourselves to words consisting uniquely of letters present in the alphabet of the corresponding language and the symbol “ ’ ” (apostrophe). This is meant as a conservative approach in order to minimize the influence of foreign words, numbers (e.g. prices), or scanning problems which are present in the raw data. In the second step, when constructing yearly data $y(t)$, i.e., words present in books published in year t , we include only those words in the database $y(t)$, which appear more than 40 times in that particular year. In the same way, for the cumulative data $Y(t)$ we include only those words, which appeared more than 40 times until time t . In this way we avoid a possible bias due to the filtering applied in the construction of the raw data (words had to appear more than 40 times in all times in order to be included in the database [1]). As an example of possible bias, in case we had not applied this filter, take two words (called '1' and '2') with $N_1(t) = N_2(t) = 21$ occurrences in year t . If now $\forall t' \neq t : N_1(t') = 0$ and $\exists t'' \neq t : N_2(t'') > 20$, word '2' would be present in the raw data whereas word '1' would be not. As a result we would only include word '2' in the yearly database $y(t)$. With our additional filter neither word '1' nor word '2' appears in the yearly database $y(t)$.

In Fig. S1 we show the resulting database size for the yearly data $y(t)$ and the cumulative data $Y(t) = \sum_{t_0}^t y(t)$ in terms of tokens and types for English, French, Spanish, German, and Russian. In this context type refers to the number of distinct words, whereas token refers to the total number of words.

For the yearly database $y(t)$ we use data in the period $t \in [1805, 2000]$, because as already indicated in [1], the database composition may have changed in a noncontinuous way at $t \approx 1800$. This claim is supported in Fig. S2, where we calculate Kendall's rank correlation coefficient [2] $\tau[y(t), y(t')]$ between the common types of the database $y(t)$ and $y(t')$ for $1500 \leq t \leq t' \leq 2000$ as

$$\tau[y(t), y(t')] = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}, \quad (\text{S1})$$

where n is the total number of common elements, n_c the number of concordant, and n_d the number of discordant pairs between the two databases with respect to the ranking of frequencies. Clearly, at $t = 1800$ a noncontinuous change in τ can be identified, from which we conclude that database composition is dramatically different in the years before and after $t = 1800$. In order not to be affected by this change the yearly data $y(t)$ is only considered in the period $t \in [1805, 2000]$. However, in order to take advantage of the full size of the database, the cumulative data $Y(t)$ is constructed taking into account all the years prior to $t = 1805$.

II. MAXIMUM LIKELIHOOD ESTIMATION

A. Theory

In this section we give account of the distributions proposed for fitting the rank-frequency distribution and present the details of the Maximum likelihood estimation procedure. The procedures are standard [3], but here we fit directly the rank frequency distribution originally proposed by Zipf [4] instead of the word frequency distribution considered in Ref. [5].

In Tab. S1 the proposed descriptive models used to fit the rank-frequency distribution are presented. The notation $F(r; \Omega)$ means that the distribution F depends on the rank r , and Ω is the set of parameters. The normalization constant $C = C(\Omega)$ is a function of the respective parameters and fixed by $\sum_{r=1}^{\infty} F(r; \Omega) = 1$. In practice, this is calculated with the Euler-Maclaurin formula available in the package `mpmath` [6].

The parameters of each distribution are estimated numerically by minimizing the negative of the log-Likelihood

$$\Omega^* = \arg \min_{\Omega} \mathcal{L}'(\Omega), \quad (\text{S2})$$

i	distribution	$F(r; \Omega)$	set of parameters Ω
1	Power-Law	$Cr^{-\gamma}$	γ
2	Shifted Power-Law	$C(r+b)^{-\gamma}$	γ, b
3	Power-Law with Exponential cutoff (beginning)	$C \exp(-b/r) r^{-\gamma}$	γ, b
4	Power-Law with Exponential cutoff (tail)	$C \exp(-br) r^{-\gamma}$	γ, b
5	Log-normal	$Cr^{-1} \exp(-\frac{1}{2}(\ln r - \mu)^2 / \sigma^2)$	μ, σ
6	Weibull	$Cr^{\gamma-1} \exp(-br^{-\gamma})$	γ, b
7	Double Power-Law	$C \begin{cases} r^{-1}, & r \leq b \\ b^{\gamma-1} r^{-\gamma} & r > b, \end{cases}$	γ, b

TABLE S1. Proposed models to fit rank-frequency distributions.

where

$$\mathcal{L}'(\Omega) = -\ln \mathcal{L} = -\sum_{i=1}^M \ln F(r(i); \Omega). \quad (\text{S3})$$

In this expression M is the number of tokens, which implies that the sum goes over each observed token i and its corresponding rank $r(i)$. In practice, the minimization is obtained with a Nelder-Mead simplex algorithm (available in the Scipy library [7]).

The quality of the fit was evaluated quantitatively by means of a p -value obtained from a χ^2 -statistics [8]:

$$\chi^2 = \sum_{j=1}^Q = \frac{(N_j - n_j)^2}{n_j}. \quad (\text{S4})$$

Here the domain is partitioned into Q cells, such that the expected number of observations per cell $n_j \geq 5$ [9], with N_j being the actual observed number of observations in cell j . A recently proposed alternative strategy [10] involving the comparison of the Kolmogorov-Smirnow statistics of the actual empirical data with randomly generated data is computationally not feasible in this case, because it would require us to draw $\approx 10^{15}$ random numbers (p -value precision 0.01) due to the size of the database of $> 10^{11}$ tokens.

In the last step we determine which of the proposed models $i = 1 \dots R$, where R is the number different models considered, is most likely to describe the data. In order to account for the different number of fitted parameters we calculate the Akaike information criterion (AIC) [11] for each model i

$$AIC = 2\mathcal{L}'(\Omega^*) + 2K, \quad (\text{S5})$$

where K is the number of parameters estimated in the model. The model which gives the minimum value $AIC_{\min} = \min_i \{AIC_i\}$ is most likely to describe the given data. From this we can calculate the relative likelihood l_i [12]

$$l_i = \exp(-(AIC_i - AIC_{\min})/2), \quad (\text{S6})$$

which states how likely model i is to describe the data in comparison with the best model. This implies that the probability w_i that model i (out of the R models considered) describes the data is given by [12]

$$w_i = P(\text{model } i | \text{data}) = l_i / \sum_{j=1}^R l_j. \quad (\text{S7})$$

B. Results

In this section we give a detailed overview of the results obtained from fitting the models in Tab. S1 to the rank-frequency distributions for all languages considered, i.e., English, French, Spanish, German, and Russian. In Fig. S3 - S7(a+b) we plot the AIC from the models in Tab. S1 applied to yearly $y(t)$ and cumulative data $Y(t)$ of the respective language. In Fig. S3 - S7(c) we show explicitly the rank-frequency distribution of the data $Y(2000)$ and the corresponding fits of the three models that yield the best description: the double power-law ($i = 7$), the power-law with an exponential cutoff in the tail ($i = 3$), and the log-normal ($i = 5$).

For English, $i = 7$ yields the best description of the yearly data for $t \gtrsim 1950$ and for the cumulative data for $t \gtrsim 1810$. As the databases $y(1950)$ and $Y(1810)$ can be considered independent datasets and by comparing with Fig. S1(a) we conclude that the size of the database needs to exceed a certain threshold ($\approx 10^9$ tokens) in order to discriminate competing models like the $i = 3$ in the tail. This is further corroborated by looking at the inset in Fig. S3(c), where it can be seen that $i = 7$ outperforms $i = 3, 5$ especially in the description of the tail of the distribution.

For the other languages except English the AIC of the yearly data $y(t)$ favours $i = 3$. This comes with no surprise since their size is limited to $< 10^9$ tokens for all $t \in [1805, 2000]$, as can be seen in Fig. S1(a). In contrast, the cumulative data $Y(t)$ shows different results. For French and Spanish the AIC favors $i = 7$ as the size of the database grows, especially for the largest dataset $Y(2000)$. Again, this becomes clear when looking at the deviations of the fits to the real data in the inset of Fig. S4(c), S5(c), which seem to diverge for $i = 3, 5$ in the tail of the distribution. For German and Russian the AIC identifies $i = 7$ only as the second best fit for the cumulative data $Y(t)$. This is most probably due to the fact that the size of the database for those languages is still not large enough in order to discriminate a second power-law regime clearly. Additionally, for these languages the critical rank b^* , where a transition between the two power-laws occurs, is shifted towards higher values, possibly due to the different degree of inflection (see main text). This in turn implies that the fraction of tokens belonging to the power-law in the tail is much smaller than in English, which means that a larger database is needed in order to discriminate $i = 3, 5$. This claim is further supported by the insets of Fig. S6(c), S7(c), where we show that especially in the tail of the distribution $i = 7$ deviates less from the data than the competing models.

Whereas English, French, and Spanish give approximately the same values for the largest database $Y(2000)$, German and Russian show larger values for b and a different power-law exponent in the tail (see main text). The latter might point towards more subtle differences between the languages besides inflection.

III. ZIPFIAN ENSEMBLE

A. Theory

The Zipfian Ensemble (ZE) [13] is a simple approach to model the size of the vocabulary depending on the text length given the rank-frequency distribution $F(r)$, $r = 1 \dots N_{ZE}^{\max}$, where $N_{ZE}^{\max} \in [1, \infty)$ is the hypothetical (maximum) size of the vocabulary. The occurrence of each word-type with rank r is assumed to be governed independently by a Poisson process with an intensity equal to its frequency, e.g., $\lambda(r) = F(r)$, where time is measured in units of tokens M (text length). This means that the probability that this word-type occurs at least once in the interval $T_1 \in [0, M]$ is given by [14]

$$P(T_1 \leq M; r) = 1 - e^{-F(r)M}. \quad (S8)$$

From this we can calculate the vocabulary size $N(M)$ by summing over all word-types, which gives the expected (average over realizations) number of words (out of N_{ZE}^{\max} different words in total) that appeared at least once up until time M :

$$N(M) = \sum_{r=1}^{N_{ZE}^{\max}} \left[1 - e^{-F(r)M} \right]. \quad (S9)$$

The variance of the ZE over the different realizations indicates the expected fluctuation around $N(M)$ in Eq. (S9) and is given by [13]:

$$\mathbb{V}[N(M)] = N(2M) - N(M). \quad (S10)$$

Although similar, this framework differs from the usual 'bag-of-words' [15] in the sense that i) the expected time of occurrence of a word need not to be an integer and ii) two words can in principle occur at the same time due to the independence of the Poisson processes. This in turn limits the interpretation of the ZE as a model for the creation of a text token by token. However, it allows for an analytic treatment and the continuous time approximation becomes better in the limit of large databases.

B. ZE in the double power-law

In this section we want to show that a double power-law in the rank frequency distribution (Eq. (1) main text) can lead to the double scaling in the vocabulary growth (Eq. (2) in main text) in the framework of the ZE.

First, we generalize the ZE to cases where words have to appear at least n times before they are considered part of the vocabulary. The introduction of a threshold n means that instead of looking at the probability for the time until its first occurrence T_1 , one considers T_n , the time it takes until the word occurs n times and Eq. (S8) becomes

$$P(T_n \leq M; r) = 1 - \sum_{j=0}^{n-1} \frac{(F(r)M)^j}{j!} e^{-F(r)M}. \quad (\text{S11})$$

From this, Eq. (S9) can be directly extended to

$$N^{(n)}(M) = \sum_{r=1}^{N_{\text{ZE}}^{\text{max}}} \left[1 - \sum_{j=0}^{n-1} \frac{(F(r)M)^j}{j!} e^{-F(r)M} \right]. \quad (\text{S12})$$

In the next step we consider the limit $n \gg 1$. As the stochastic variable T_n is the sum of n times the stochastic variable T_1 , which is distributed according to Eq. (S8), one can conclude that by means of the central limit theorem it follows that $P(T_n = M/n; r)$ will approach a Gaussian with vanishing variance, such that by rescaling $M \mapsto M/n$ Eq. (S11) asymptotically becomes

$$\lim_{n \rightarrow \infty} P(T_n \leq M/n; r) = \Theta(M/n - \tau(r)), \quad (\text{S13})$$

where $\tau(r) = 1/F(r)$ is the inverse of the frequency $F(r)$ of the particular word-type and $\Theta(x)$ is the Heaviside step function. For the vocabulary growth this yields

$$\lim_{n \rightarrow \infty} N^{(n)}(M/n) = \sum_{r=1}^{N_{\text{ZE}}^{\text{max}}} [1 - \Theta(M/n - \tau(r))]. \quad (\text{S14})$$

Thus we obtain a direct relationship between the rank-frequency distribution and the vocabulary growth

$$\lim_{n \rightarrow \infty} N^{(n)}(\tilde{M} = 1/F(r)) = r, \quad (\text{S15})$$

where $\tilde{M} = M/n$.

In Fig. S8 we show the $N^{(n)}(\tilde{M})$ curve obtained from the ZE for the double power-law (Eq.(1) main text) with parameters $\gamma^* = 1.77$ and $b^* = 7873$ for different thresholds n . One can see that the growth curves for $n > 8$ are almost indistinguishable from the asymptotic solution Eq. (S15), which can be attributed to the fast convergence implied by the central limit theorem.

From these observation we conclude that Eq. (S15) is already a good approximation for $n \gg 1$, where in practice this can mean $n > 10$. As a result we obtain Eq. (2) from the main text. This means that the increase of the threshold n leads to a reduction of the fluctuations of the growth curve of the vocabulary and can be explained as a result of a simple stochastic process. In Fig. S9 we show that this claim holds when applied to real texts of the size of single books, as well as for a collection of several million books, as in Fig. S10.

IV. NUMERICAL SIMULATION OF STOCHASTIC MODEL

In this section we show the results of the direct numerical simulation of the model proposed in Sec. III, main text.

A. Parameters and initialization

In order to simulate the model, apart from fixing a number of parameters ($N_c^{\text{max}}, \alpha, p_c^0$), we need to prescribe how the model is initialized, e.g., what is the initial probability of using a new word p_{new}^0 and how many word types exist at the first iteration of the model. Concerning the parameters, the initial probability of choosing a core-word is set to $p_c^0 = 0.99$, such that $1 - p_c^0 \ll 1$ (see main text) and the two other parameters are fixed by the fitting parameters ($N_c^{\text{max}} = b^* = 7873$, $\alpha = \gamma^* - 1 = 0.77$ in English, see main text). Concerning the initialization of the model, an important point that needs to be taken into account is that we are interested in retrieving the Heaps' plot obtained after re-scaling the number of tokens M by the threshold n as $\tilde{M} = M/n$, see Fig. S10 (for simplicity and computational efficiency in our simulations we choose $n = 1$). This implies that the first word type of our model

should on average appear not at the first token but instead approximately at $\tilde{M} \approx 1/F(1)$ (where $F(1)$ is the frequency of the most frequent word). In view of this requirement, we set $p_{\text{new}}^0 = C = F_{dp}(1) = 0.0922$ (for English, see main text) and we start with an empty list of word types (the tokens used before the appearance of the first word type are counted but not attributed to any word type). The simulations were done with a maximum number of $M = 10^9$ steps in units of tokens, a restriction imposed by the computational effort required. The reported results were obtained as the average of 100 realizations of the model.

B. Heaps' plot

In order to be able to compare the results with the google-ngram data, where a natural threshold of $n = 41$ is imposed (see SI-Sec. I), we incorporate the threshold n by using the rescaled coordinate $\tilde{M} = M/n$, as motivated and discussed in SI-Sec. III B. In Fig. S11 we show the expected vocabulary growth $N(\tilde{M})$. We can clearly see that the two scaling regimes of Eq. (2), main text, are recovered from our model. Deviations to the data are within 50% over as much as 7 orders of magnitude. The poorer agreement for large \tilde{M} can be attributed to a slight overestimation by our model of the point of transition between the two scaling regimes. This could be addressed by modifying our model (e.g. modifying our simple choice of p_c in Eq. (3) main text) so that the decay in p_{new} and the transition to the second scaling occurs already for shorter \tilde{M} . For even larger \tilde{M} we do not have data for our model due to computational limitations. However, based on our asymptotic calculations in Sec. III, main text, we expect that the observed agreement will extend over the entire range of available data.

C. Zipf's plot

In the analysis of the Zipf's plot $F(r)$ obtained by our model it is important to take into account that the Yule's type processes (already used words are drawn proportional to their previous occurrences) give a disproportionately large weight to the first word types used in the simulation. This happens because in the beginning of the simulation there are only a few existing word types into which all drawn tokens are attributed. Figure S12(a) illustrates this effect and shows that it is inversely proportional to p_{new}^0 , which sets the time-scale for the appearance of new word types in the beginning of the simulation. This artifact can be easily addressed by excluding a few word types of smallest rank and re-normalizing the remaining distribution, as shown in Fig. S12(b-d). Alternatively, one can modify the preferential attachment part of the model (right-most branch in Fig. 3, main text) so that the very first used word-types follow a different rule and have a vanishing probability of usage for large M (notice that this would not modify the Heap's plot). For the case of English, Fig. S13 shows that the removal of only one word type ($r = 1$) is sufficient in order to obtain a good agreement with data (less than 50% of deviation over 7 orders of magnitude). As discussed in the case of Heaps' plot above, the transition to the second scaling appears slightly shifted in comparison to the data.

-
- [1] J.-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden, *Quantitative analysis of culture using millions of digitized books*, Science **331**, 176 (2011).
 - [2] M.G. Kendall, *Rank Correlation Methods* (Griffin, 1970), 4th ed.
 - [3] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes* (Cambridge University Press, 2007), 3rd ed.
 - [4] George Kingsley Zipf, *The Psycho-Biology of Language* (Routledge, 1936).
 - [5] M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemp. Phys. **46**, 323 (2005).
 - [6] F. Johansson et al., *mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.14)* (2010), <http://code.google.com/p/mpmath/>.
 - [7] E. Jones, T. Oliphant, P. Peterson et al., *SciPy: Open source scientific tools for Python*, (2001–), <http://www.scipy.org/>.
 - [8] R.B. D'Agostino, *Goodness-of-Fit-Techniques* (Taylor Francis, 1986).
 - [9] J.R. Taylor, *An Introduction to Error Analysis* (University Science Books, 1997).
 - [10] A. Clauset, C.R. Shalizi, M.E.J. Newman, *Power-Law Distributions in Empirical Data*, SIAM Rev. **51**, 661 (2009).
 - [11] H. Akaike, *A new look at the statistical model identification*, IEEE T. Automat. Contr. **19**, 716 (1974).
 - [12] K.P. Burnham, and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, 2002), 2nd ed.
 - [13] I. Eliazar, *The growth statistics of Zipfian ensembles: Beyond Heaps' law*, Physica A **390**, 3189 (2011).

- [14] D.R. Cox, and H.D. Miller, *The Theory of Stochastic Processes* (Chapman Hall, 1977).
- [15] K.W. Church, and W.A. Gale, *Poisson Mixtures*, Nat. Lang. Eng. **1**, 163 (1995).

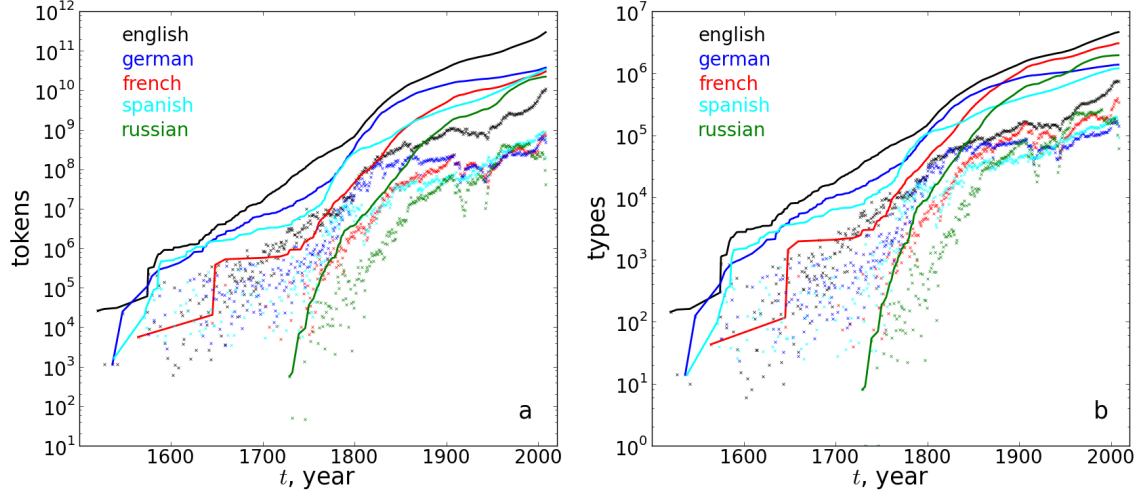


FIG. S1. Size of the database after filtering. a) Number of tokens for yearly data $y(t)$ (x-symbols) and cumulative data $Y(t)$ (line). b) Number of types for yearly data $y(t)$ (x-symbols) and cumulative data $Y(t)$ (line). Each language is marked by a different color.

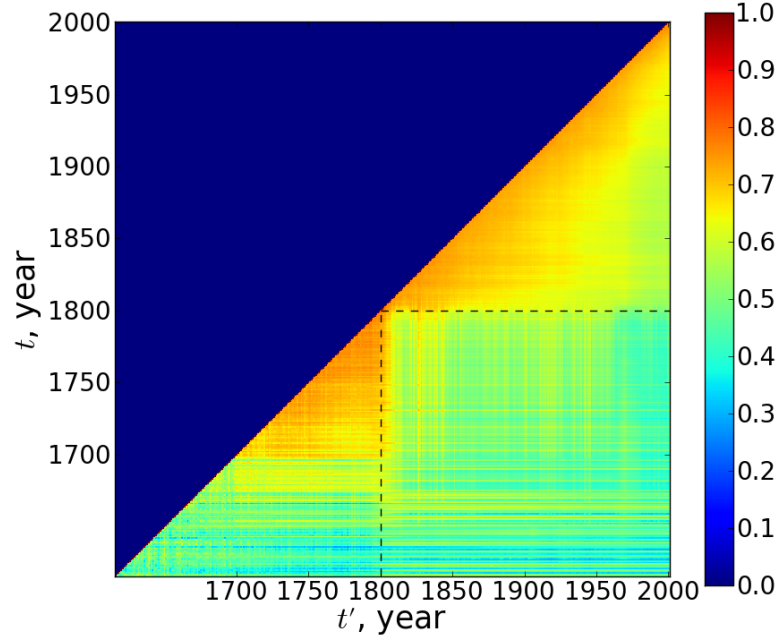


FIG. S2. Correlation between data in different years for English. Kendall's rank correlation Eq. (S1) between yearly data $y(t)$, $y(t')$ for $t, t' \in [1500, 2000]$ with $t \leq t'$. The dashed lines show $t = 1800$ and $t' = 1800$ where a noncontinuous change in the correlation occurs.

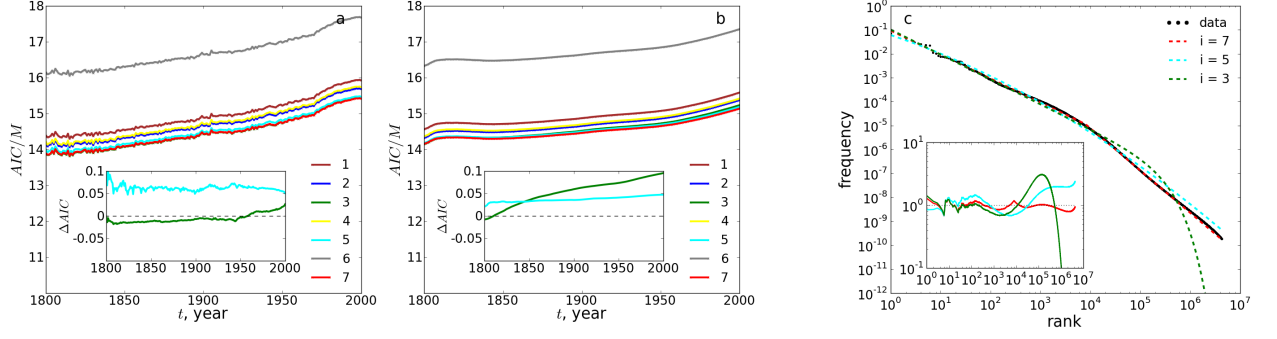


FIG. S3. Discrimination between different models with AIC for English. Value of the AIC for a) yearly data $y(t)$ b) cumulative data $Y(t)$. The inset shows the difference $\Delta AIC = AIC_i/M - AIC_7/M$, $i = 1..6$ meaning that if $\Delta AIC > 0$ the double power-law is the most likely model among the proposed describing the data. Numbers refer to the indices of the model in Tab. S1. c) rank-frequency plot for $Y(2000)$ and the fits of the three best models. The inset shows the ratio $F_{data}(r)/F_{fit}(r)$.

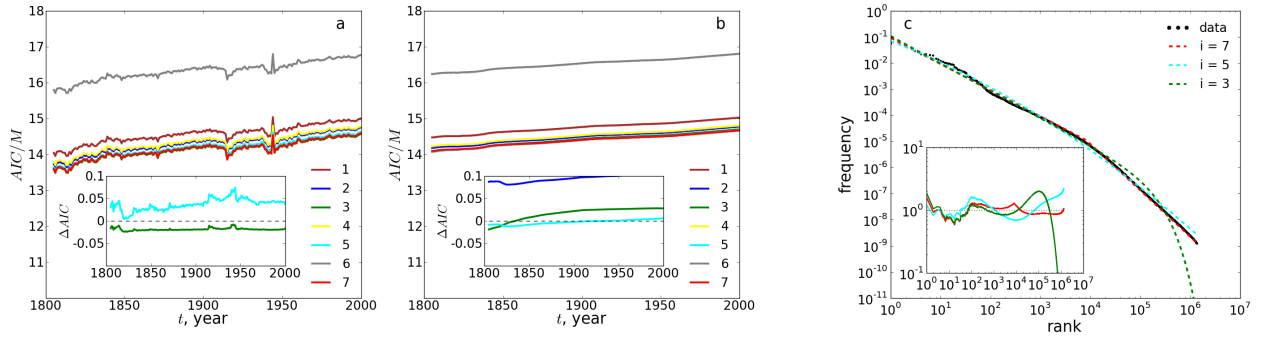


FIG. S4. Same as in Fig. S3 for French.

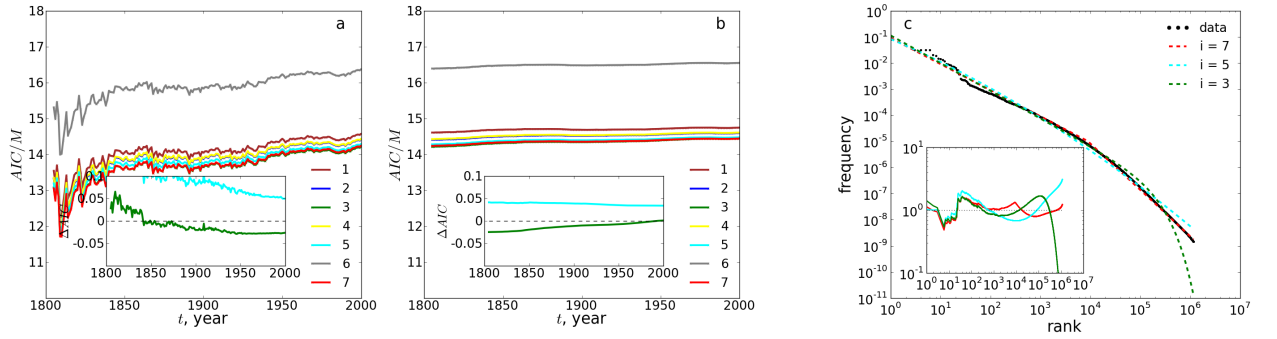


FIG. S5. Same as in Fig. S3 for Spanish.

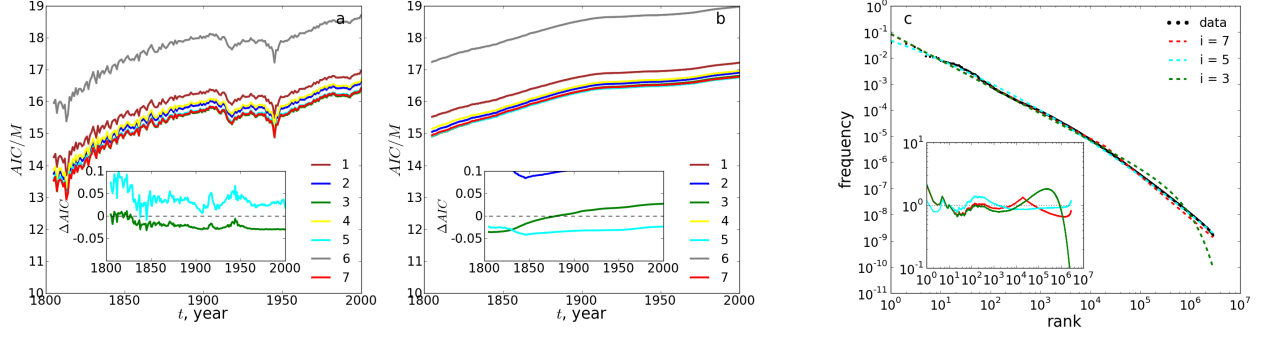


FIG. S6. Same as in Fig. S3 for German.

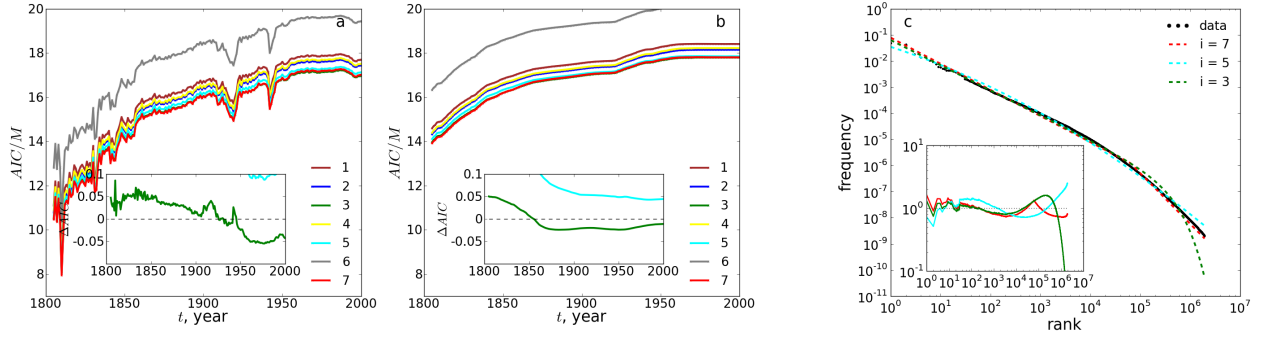
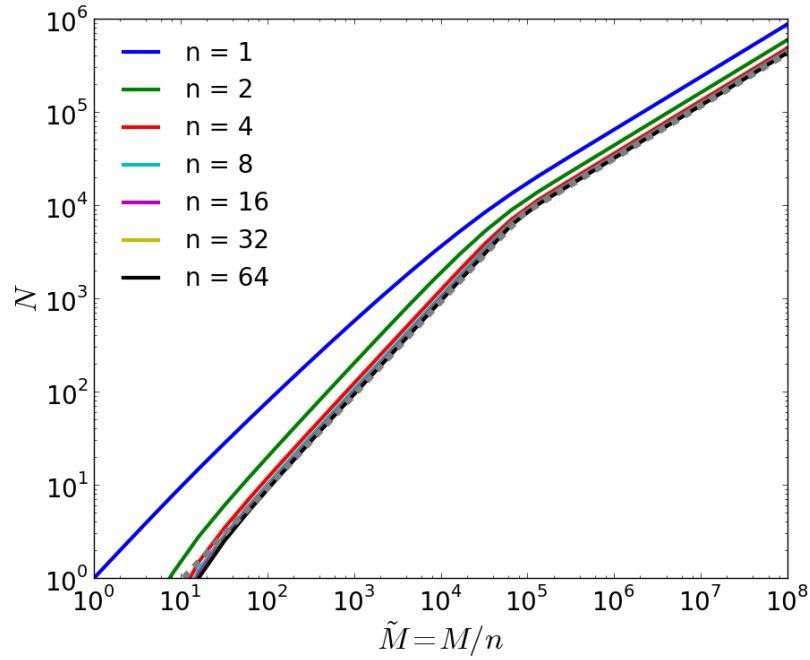


FIG. S7. Same as in Fig. S3 for Russian.

FIG. S8. Influence of threshold n on size of vocabulary for the ZE. Growth curves $N(\tilde{M} = M/n)$ obtained from ZE for double power-law (Eq.(1) main text) with parameters $\gamma^* = 1.77$, $b^* = 7873$ with different thresholds n . The dashed curve shows the asymptotic solution Eq. (S15).

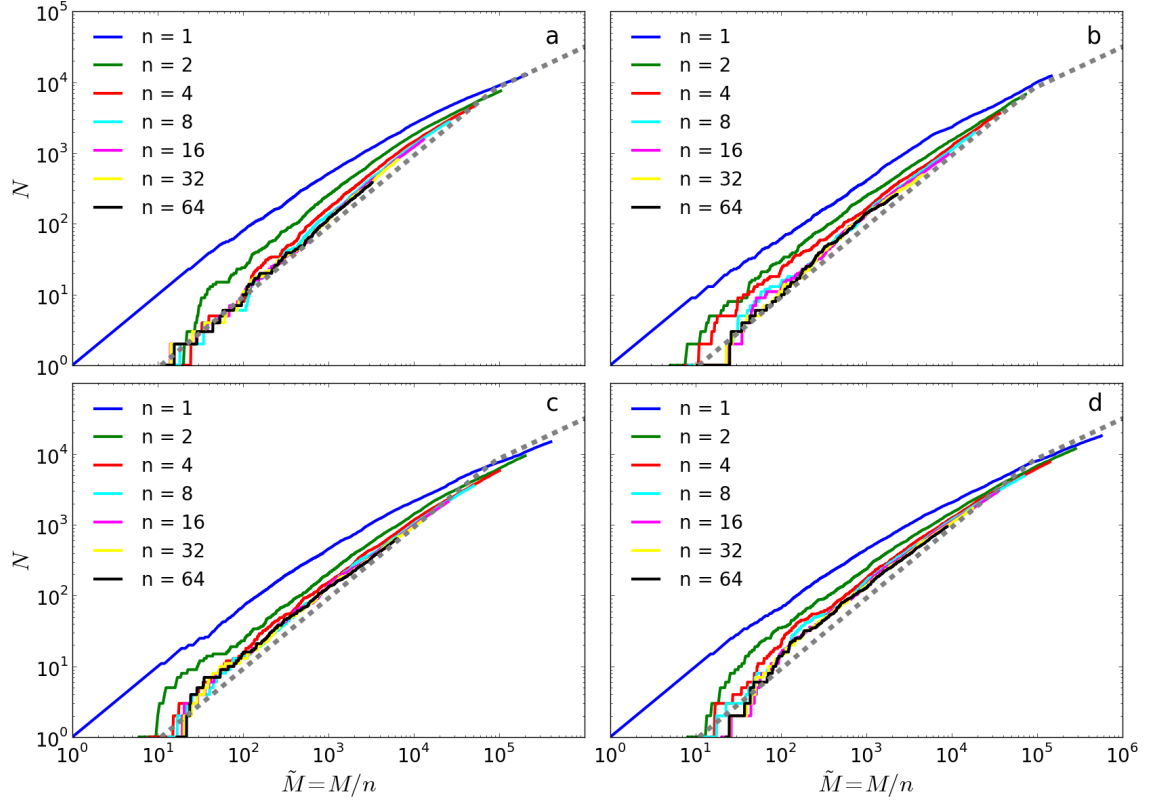


FIG. S9. Influence of threshold n on size of vocabulary for single books. Growth curves $N(\tilde{M} = M/n)$ obtained from 4 different books with different thresholds n . a) Charles Darwin: “The Voyage of the Beagle” b) Mark Twain: “Life on the Mississippi” c) Miguel de Cervantes Saavedra: “Don Quixote”, translated by John Ormsby d) Leo Tolstoy: “War and Peace”, translated by Louise and Aylmer Maude. All texts were retrieved from the Project Gutenberg (www.gutenberg.org) on 21.09.2010. The dashed curve shows the asymptotic solution Eq. (S15) of the ZE assuming a double power-law (Eq.(1) main text) with parameters $\gamma^* = 1.77$, $b^* = 7873$.

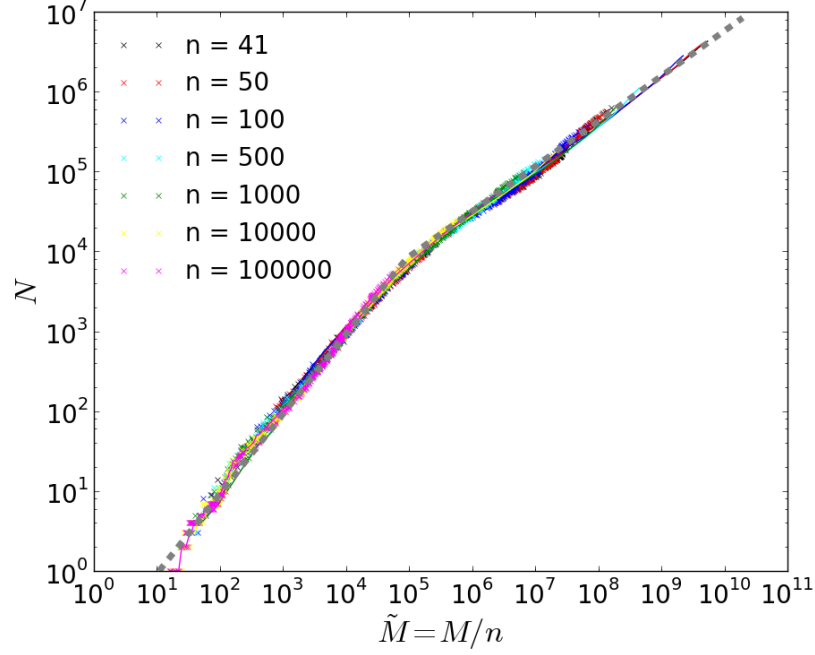


FIG. S10. Influence of threshold n on size of vocabulary for the English google-ngram database. Growth curves $N(\tilde{M} = M/n)$ obtained from yearly data $y(t)$ (x-symbol) and cumulative data $Y(t)$ (line) for different values of the threshold n with $n \in [41, 10^5]$ marked by different colors. The dashed curve shows the asymptotic solution Eq. (S15) of the ZE assuming a double power-law (Eq.(1) main text) with parameters $\gamma^* = 1.77$, $b^* = 7873$.

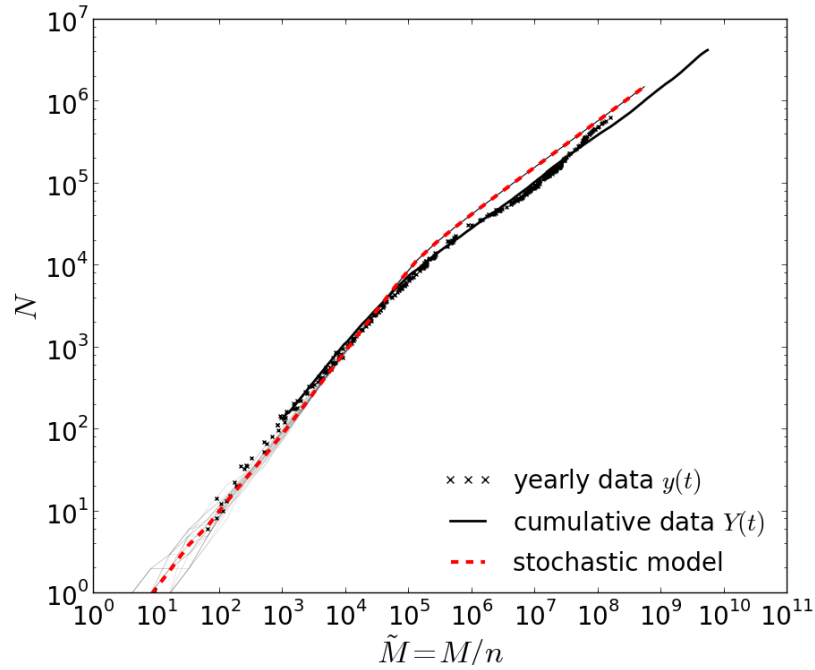


FIG. S11. Vocabulary growth, $N(\tilde{M})$, from the numerical simulation of our stochastic model (Heaps' plot). Number of word-types as a function of word-tokens of the English database for yearly (x-symbols) database, cumulative (solid) database, and the expectation from our stochastic model (dashed). Single realizations of the stochastic process are shown in thin/gray (solid). Each realization is calculated for an imaginary text of $\tilde{M} = 10^9$ tokens.

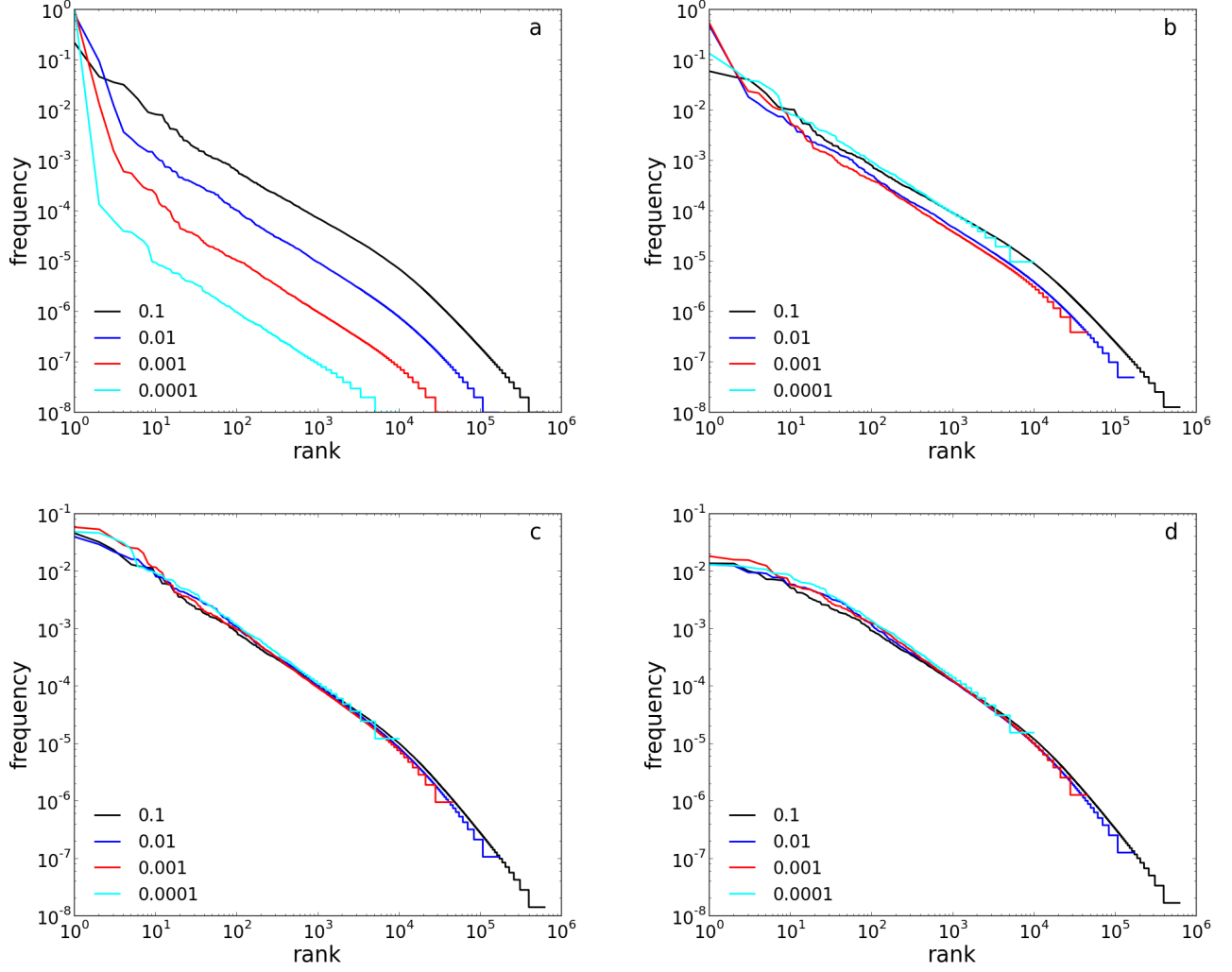


FIG. S12. Influence of the first word types on the rank-frequency distribution of our model. Rank-frequency distribution $F(r)$ from our numerical simulation with different values for $p_{\text{new}}^0 \in \{0.1, 0.01, 0.001, 0.0001\}$ after filtering the k most frequent types, where a) $k=0$, b) $k=1$, c) $k=3$, and d) $k=10$. In this context, filtering means, that i) we neglect all tokens associated with ranks $r = 1 \dots k$; ii) the rank of all remaining types is lowered by k , e.g., the rank of the $k+1$ -th most frequent type becomes $r=1$; and iii) the distribution is renormalized such that $\sum_{r=1}^{N-k} F(r) = 1$, where N is the number of types before the filtering.

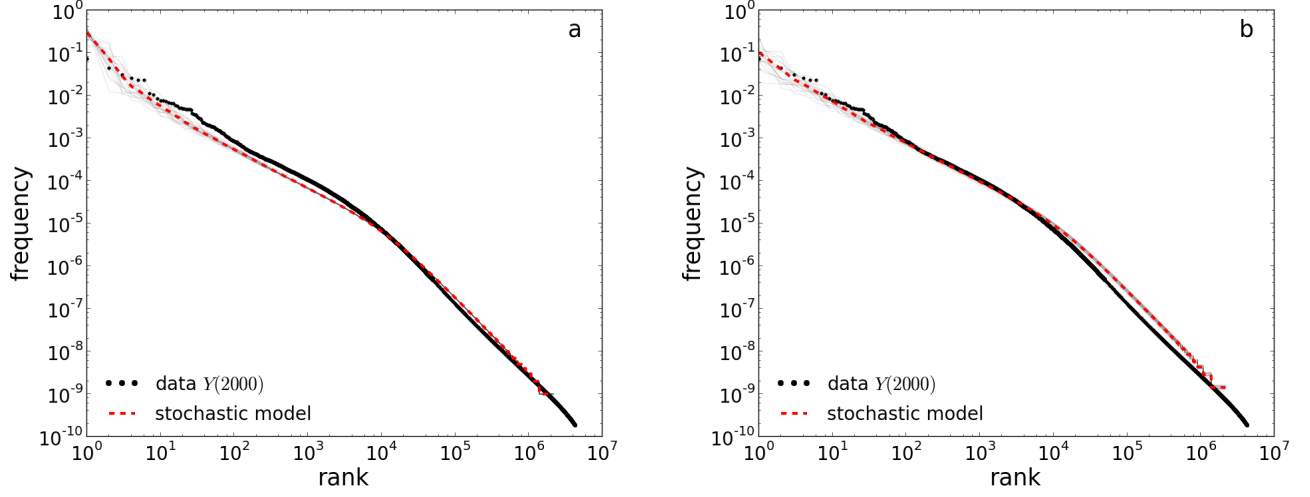


FIG. S13. Rank frequency distribution, $F(r)$, from the numerical simulation of our stochastic model (Zipf's plot). Rank-frequency distribution for the English database $Y(2000)$ (solid) and the expectation from our stochastic model (dashed), where a) shows the unfiltered result, and b) shows the distribution after filtering the type which has rank $r = 1$ in a). Single realizations of the stochastic process are shown in thin/gray (solid). Each realization is calculated for an imaginary text of $\tilde{M} = 10^9$ tokens.